EMERGENT SYMBOLIC ORGANIZATION IN SPARSE BINARY VISUAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Symbolic organization is often associated with language, but whether it can emerge within visual recognition without linguistic supervision remains unclear. By imposing a sparse binary bottleneck on ResNet-50, we represent each image as a set of active bits from a shared vocabulary of 2048 binary visual units and show that symbolic organization can emerge from image-level supervision alone. The binary representation activates only 24.63 bits per image on average, corresponding to 98.8% sparsity, while retaining 78.54% Top-1 accuracy on ImageNet. This shared vocabulary exhibits systematic reuse, category-structured combinations, and a broad Zipf-like regime in its rank–frequency curve, together with functionally differentiated contributions to recognition. To probe the semantics of individual bits, we map the 2048 bits to learnable embeddings and use them to condition a frozen Stable Diffusion 1.5 model. The experiments show that complete active-bit sets generate semantically recognizable visual content. Some individual bits produce stable semantic effects, including concepts beyond the ImageNet label taxonomy, and independently probed bits can combine into coherent configurations beyond observed ImageNet compositions. At least a subset of the learned visual bits therefore exhibits multiple properties characteristic of symbolic units, providing an empirical setting for studying how symbolic organization can emerge in perceptual neural systems. Code will be made publicly available.

1 INTRODUCTION

Language organizes complex experience through a finite inventory of discrete symbols that can be shared across instances, reused systematically, and combined into larger structures (Fodor & Pylyshyn, 1988; Newell & Simon, 2007). Through reuse and composition, such a shared inventory can support a much richer space of representations than its size alone would suggest (Lake et al., 2016; Pelletier, 2017; Nam et al., 2023). However, whether these symbolic organizational principles are unique to language, or can also emerge within visual recognition without linguistic supervision, remains an open question in artificial intelligence and cognitive science (Harnad, 1990).

Several studies have explored related aspects of discrete representation and symbolic organization. Sparse and discrete representation learning has shown how neural codes can be constrained into compact or discrete units (Olshausen & Field, 1996; Van Den Oord et al., 2017; Yamazaki et al., 2022; Yuan & Agaian, 2023), while interpretability studies investigate the properties and functional roles of individual learned features (Wang et al., 2020; Zhang et al., 2021; Lee et al., 2025). Emergent-communication studies examine how shared discrete protocols arise under interaction (Lazaridou & Baroni, 2020; Boldt & Mortensen, 2024; Deng et al., 2025), and vision-language models align visual representations with language-grounded semantics through large-scale image-text supervision (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023). These directions have established important methods for studying discrete representation and semantic organization, but leave open whether a visual recognition system itself can organize discrete internal units into a shared, reusable vocabulary using only image-level supervision.

We investigate this question by imposing a binary $\{0, 1\}$ constraint on the 2048-dimensional feature vector extracted by ResNet-50 (He et al., 2016) and using the resulting discrete representation as the input to the classification head. Each image is thereby represented as a sparse set of active bits drawn from a shared inventory of 2048 discrete visual units. The recognition model is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

trained only with image-level class labels, without language supervision, concept annotations, or an explicit symbolic objective. This simple constraint produces a highly sparse representation that nevertheless retains substantial recognition capability: the model achieves 78.54% ImageNet Top-1 accuracy while activating only 24.63 bits per image on average, corresponding to 98.8% sparsity. More importantly, the discrete bits develop non-trivial statistical organization over the shared inventory. Their rank–frequency curve exhibits a pronounced head–core–tail distribution with a broad Zipf-like regime reminiscent of natural-language vocabularies (Cancho & Solé, 2003; Piantadosi, 2014; Zipf, 2016), and same-class images repeatedly reuse coherent combinations of bits. Counterfactual deletion experiments further reveal differentiated functional roles, with targeted removal of analytically important bits producing substantially larger recognition changes than random removal.

To further probe what visual information is carried by these discrete bits, we map the active-bit representation of each image to learnable embeddings and use them as the conditioning signal for a frozen text-to-image model, Stable Diffusion 1.5 (Rombach et al., 2022). This allows the learned discrete representation to be evaluated through its ability to control a pretrained image generator. Experiments show that complete active-bit representations retain sufficient information to generate recognizable visual content, while some individual bits produce stable visual effects across random seeds. Some bits are associated with visual concepts beyond the supervised ImageNet label taxonomy, such as statue-like and cream-cake-like patterns. Moreover, independently activated bits can be jointly invoked, allowing their visual effects to combine into coherent novel configurations beyond the composition distribution of the training data. These results show that some individual bits can exert stable, independently addressable semantic effects and can participate in novel combinations.

More broadly, these findings suggest that symbol-like organization can emerge within visual recognition without linguistic supervision. A shared inventory of reusable discrete units may provide a more primitive organizational layer, from which richer forms of reference and composition, and eventually sequence, grammar, and communication, may develop. This perspective opens a broader avenue for studying how linguistic systems may build upon more primitive forms of symbolic organization in perceptual representations.

2 Related Work

Research on discrete and sparse representations examines how neural systems can encode information under constraints on their internal codes. Sparse autoencoder methods show how neural representations can be decomposed into sparse learned features (Rajamanoharan et al., 2024; Gao et al., 2025; Templeton et al., 2026), while quantized representation learning introduces explicit discrete latent codes for neural models (Yu et al., 2024; Mentzer et al., 2024; Sun et al., 2024). More extreme forms of discretization have been explored in binary neural networks (Hubara et al., 2016; Ma et al., 2024; Colombo et al., 2026) and spiking neural networks (Shen et al., 2025; Ma et al., 2026a;b), which constrain weights, activations, or computational states into binary representations. These works demonstrate that neural systems can preserve task-relevant information under strong discreteness or sparsity constraints. However, they focus primarily on representation efficiency, optimization, or computation, rather than on whether discrete units can self-organize into shared symbolic structures with systematic reuse and composition.

Sparse distributed representations (SDRs) (Ahmad & Hawkins, 2015) provide a mathematical framework for understanding the properties of sparse binary codes. SDR theory shows that distributed binary representations can achieve high representational capacity and robustness through sparse activation patterns. Because each item is represented by a sparse subset of a shared neural population, such representations can encode rich patterns through combinatorial reuse. These properties suggest that sparse binary codes provide a potential substrate for structured information organization. However, SDR studies primarily characterize the theoretical properties of such representations; they typically do not address how such representations emerge in learned systems or whether they can acquire semantic and functional organization through perceptual learning.

Interpretability research investigates how learned neural representations relate to human-interpretable concepts and functional behaviors. Network dissection and concept-based approaches associate individual neural features with semantic concepts (Bau et al., 2017; Oikarinen et al., 2023; Kowal et al., 2024), while causal interpretability methods analyze their functional roles through interventions (Kim et al., 2018; Shin et al., 2023; Geiger et al., 2024). These studies reveal that neural

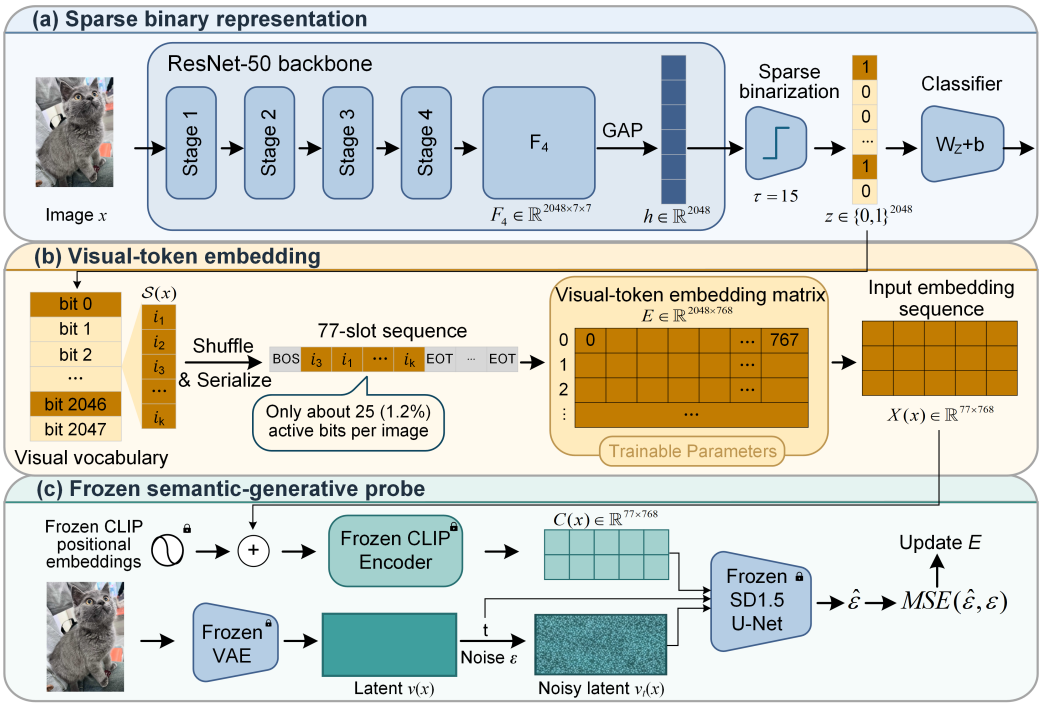


Figure 1: Overview of the proposed visual-token framework. (a) ResNet-50 maps each image to a sparse binary representation for classification. (b) The active-bit set is randomly shuffled and mapped to visual-token embeddings through a trainable embedding matrix. (c) The resulting embeddings condition a frozen Stable Diffusion 1.5 model through its CLIP encoder, while only the visual-token embedding matrix is optimized. The learned visual tokens are subsequently compared with the original CLIP vocabulary to obtain a lexical readout in the language-grounded CLIP space.

networks contain meaningful internal structures beyond their explicit training objectives. However, existing interpretability methods primarily analyze semantic properties of continuous representations or rely on external concepts to interpret learned features. Whether visual learning itself can organize discrete units with interpretable semantic meanings remains largely unexplored.

Another line of research studies how shared symbolic structures emerge through interaction and communication. Emergent communication shows that interacting agents can develop shared discrete protocols under cooperative objectives (Mu & Goodman, 2021; Rita et al., 2022a;b), demonstrating how discrete units can acquire shared meanings through interaction. However, these structures arise under explicit interaction and communication objectives rather than visual recognition alone. Whether visual recognition alone can self-organize a shared vocabulary from image-level supervision remains largely unexplored.

3 Method

3.1 Sparse Binary Visual Representation

A reasonable functional abstraction of visual recognition is that a sparse subset of elements identified from a shared visual vocabulary can be combined to distinguish visual categories. We instantiate this abstraction by imposing a sparse binary bottleneck on the 2048-dimensional pre-classification features of a ResNet-50 backbone, as illustrated in fig. 1(a).

Let x be an image and y its class label. After x passes through the four stages of ResNet-50, the final feature map $\mathbf{F}_4(x)$ is globally average-pooled to produce a feature vector

$$\mathbf{h}(x) = \mathrm{GAP}(\mathbf{F}_4(x)) \in \mathbb{R}^d, \qquad d = 2048. \tag{1}$$

Each component $h_i(x)$ is binarized as

$$z_i(x) = \mathbf{1}_{\{h_i(x) \geq \tau\}}, \qquad i \in \{0, \ldots, 2047\}, \qquad \mathbf{z}(x) \in \{0,1\}^d, \qquad \tau = 15. \tag{2}$$

Here, $\mathbf{1}_{\{\cdot\}}$ is the indicator function. A value of $z_i(x) = 1$ marks bit i as active for image x, whereas $z_i(x) = 0$ marks it as inactive. The threshold τ controls sparsity because a larger value requires a stronger feature response for activation. This operation converts the continuous feature vector $\mathbf{h}(x)$ into the sparse binary representation $\mathbf{z}(x)$.

Only $\mathbf{z}(x)$ is supplied to the classifier. The classifier produces class logits and probabilities as

$$\mathbf{o}(x) = W\mathbf{z}(x) + \mathbf{b}, \qquad \mathbf{p}(x) = \mathrm{softmax}(\mathbf{o}(x)). \tag{3}$$

For an image-label pair (x, y), the classification objective is

$$\mathcal{L}_{\mathrm{cls}}(x, y) = -\log p_y(x) \tag{4}$$

where $p_y(x)$ is the predicted probability assigned to the ground-truth class. All ResNet-50 parameters are fine-tuned end-to-end from pretrained weights using only this image-level classification objective. No language supervision, concept annotations, or special auxiliary regularizers are used.

The hard threshold in eq. (2) is non-differentiable. During backpropagation, a sigmoid straight-through estimator (SigmoidSTE) provides the surrogate gradient

$$\frac{\partial z_i(x)}{\partial h_i(x)} \coloneqq \alpha\, \sigma(\alpha(h_i(x) - \tau))\,[1 - \sigma(\alpha(h_i(x) - \tau))], \qquad \alpha = 4. \tag{5}$$

where $\sigma(u) = 1/(1 + \exp(-u))$ is the sigmoid function. The parameter α controls the sharpness of the surrogate transition around τ. Larger values make the surrogate closer to a hard threshold, whereas smaller values provide a smoother gradient. See Appendix for further details.

The $d = 2048$ bit indices therefore define a visual vocabulary shared across images. For each image x, the binary representation specifies the unordered active-bit set $\mathcal{S}(x)$:

$$\mathcal{S}(x) = \{i \in \{0, \ldots, 2047\} \mid z_i(x) = 1\}, \qquad K(x) = |\mathcal{S}(x)|. \tag{6}$$

Because no number constraint is imposed, the active-set cardinality $K(x)$ varies across images.

At the classifier level, the binary representation obeys an explicit additive composition rule. For class c, its logit is $o_c(x) = b_c + \sum_{i=0}^{d-1} W_{c,i} z_i(x)$. Thus, bit i contributes $W_{c,i} z_i(x)$ to the class-c logit. The same linear form defines an exact interventional effect at the classifier level. Viewing the logit as a function of $\mathbf{z}$, let $\mathbf{z}_{-i}$ denote all bits except z_i. Holding $\mathbf{z}_{-i}$ fixed, changing bit i by Δz_i produces $\Delta o_c = W_{c,i} \Delta z_i$. Hence, every bit–class pair has an analytically specified additive contribution and causal effect at the classifier level.

3.2 From Unordered Bit Sets to Visual-Token Embeddings

The trained ResNet-50 maps each image to an unordered active-bit set $\mathcal{S}(x)$. As a conditioning representation, this discrete set is analogous in form to a text prompt: its elements can be serialized and mapped to the input-embedding space of a text-conditioned image generator. This construction tests whether a representation learned from image-level classification can condition an image generator to recover the visual content of the corresponding image. This process is illustrated in fig. 1(b).

Stable Diffusion 1.5 with its CLIP encoder serves as the image-generation model. The CLIP encoder has a fixed context length of $L = 77$. Because the elements of $\mathcal{S}(x)$ are unordered, they are randomly shuffled before serialization. A BOS token is placed first, and EOT tokens fill the remaining slots after the shuffled visual-bit indices. The resulting token-ID sequence $\mathbf{i}(x)$ is defined as:

$$\mathbf{i}(x) = \mathrm{concat}\left(d_{\mathrm{BOS}}, \pi_x(\mathcal{S}(x)), \underbrace{d_{\mathrm{EOT}}, \ldots, d_{\mathrm{EOT}}}_{L - K(x) - 1} \right), \qquad L = 77. \tag{7}$$

Here, π_x denotes the random shuffle operation applied to the elements of $\mathcal{S}(x)$. The BOS and EOT tokens use their original CLIP vocabulary IDs, $d_{\mathrm{BOS}} = 49406$ and $d_{\mathrm{EOT}} = 49407$, respectively. Therefore, $\mathbf{i}(x) \in \{0, \ldots, 2047, 49406, 49407\}^L$.

Each visual-bit index is assigned a learnable 768-dimensional visual-token embedding. The token-ID sequence is converted into the input embedding sequence of the CLIP encoder through:

$$E \in \mathbb{R}^{2048\times 768}, \qquad \mathbf{X}(x) = \mathcal{E}(\mathbf{i}(x)) \in \mathbb{R}^{L\times 768}. \tag{8}$$

For visual-bit indices, $\mathcal{E}(\cdot)$ retrieves the corresponding rows from the trainable matrix E, whereas BOS and EOT are mapped to their fixed embeddings from the original CLIP.

3.3 Probing Visual-Token Embeddings with Frozen Stable Diffusion

Stable Diffusion 1.5 (SD1.5) serves as a frozen semantic-generative probe for the visual-token embeddings. During training, the sequence $\mathbf{X}(x)$ derived from image x conditions the prediction of the noise added to its latent representation, thereby providing the training signal. Only the visual-token embedding matrix E is optimized, while the trained ResNet-50 feature extractor, VAE, CLIP encoder, and diffusion U-Net remain frozen. Figure 1(c) summarizes this training process.

Let T_ψ denote the frozen CLIP encoder and U_ϕ denote the frozen Stable Diffusion 1.5 U-Net. The input embedding sequence $\mathbf{X}(x)$ is added to the frozen CLIP positional embedding matrix $\mathbf{P} \in \mathbb{R}^{L\times 768}$ and processed by T_ψ to produce the context representation:

$$\mathbf{C}(x) = T_\psi(\mathbf{X}(x) + \mathbf{P}) \in \mathbb{R}^{L\times 768} \tag{9}$$

The frozen VAE encodes the target image x into a latent representation $\mathbf{v}(x)$. At a randomly sampled diffusion timestep t, Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is added to obtain the noisy latent:

$$\mathbf{v}(x) = \mathrm{VAE}_{\mathrm{enc}}(x), \qquad \mathbf{v}_t(x) = \sqrt{\bar{\alpha}_t}\,\mathbf{v}(x) + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}. \tag{10}$$

where $\bar{\alpha}_t$ is determined by the frozen diffusion scheduler. The U-Net predicts the injected noise from the noisy latent, timestep, and context representation, giving the diffusion training objective:

$$\widehat{\boldsymbol{\epsilon}} = U_\phi(\mathbf{v}_t(x),\, t,\, \mathbf{C}(x)), \qquad \mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{x,t,\boldsymbol{\epsilon},\pi_x}[\|\boldsymbol{\epsilon} - \widehat{\boldsymbol{\epsilon}}\|_2^2]. \tag{11}$$

All remaining training and sampling settings follow the original Stable Diffusion 1.5 configuration.

3.4 Lexical Alignment of Visual Tokens

The learned visual-token embedding matrix enables direct lexical comparison between the image-derived visual vocabulary and the original CLIP vocabulary because embeddings from both occupy the same 768-dimensional input-embedding space of the CLIP encoder. Let $\mathcal{V}_{\mathrm{CLIP}}$ denote the set of tokens in the original CLIP vocabulary, and let $E_{\mathrm{CLIP}} \in \mathbb{R}^{|\mathcal{V}_{\mathrm{CLIP}}|\times 768}$ denote their corresponding token-embedding matrix. Let $\overline{E}$ and $\overline{E}_{\mathrm{CLIP}}$ denote the row-wise ℓ_2-normalized versions of the visual-token embedding matrix E and the CLIP token-embedding matrix E_{CLIP}, respectively. The cosine similarities between all visual tokens and CLIP tokens are computed simultaneously as

$$\mathbf{S} = \overline{E}\,\overline{E}_{\mathrm{CLIP}}^{\mathsf{T}} \in \mathbb{R}^{2048\times|\mathcal{V}_{\mathrm{CLIP}}|}. \tag{12}$$

Here, S_{ij} denotes the cosine similarity between learned visual token i and CLIP token j. For each visual token i, the k nearest CLIP tokens are selected as

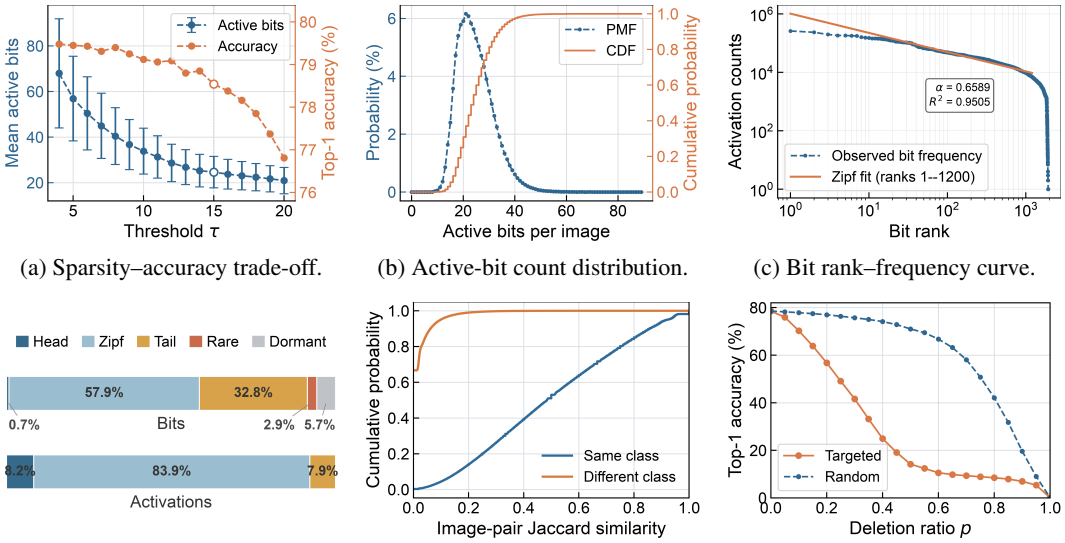
$$\mathcal{N}_k(i) = \mathrm{TopK}_{j\in\mathcal{V}_{\mathrm{CLIP}}}(S_{ij}). \tag{13}$$

The resulting $\mathcal{N}_k(i)$ contains the k nearest CLIP tokens, which serve as candidate lexical associations for the learned visual token i. This comparison makes the image-derived visual vocabulary directly interrogable through candidate natural-language concepts represented in the original CLIP space, without additional language supervision or a separate alignment network.

4 Experiments

4.1 Experimental Setup

All experiments were conducted on the ImageNet-1K (Deng et al., 2009) dataset. For recognition, a pretrained ResNet-50 model (maintainers & contributors, 2016) was fine-tuned end-to-end with image-level classification supervision using SGD with momentum 0.9, weight decay 2×10^{-5}, a



(a) Sparsity–accuracy trade-off. (b) Active-bit count distribution. (c) Bit rank–frequency curve.

(d) Visual-vocabulary usage strata. (e) Class-wise active-set overlap. (f) Random vs. targeted deletion.

Figure 2: Emergent organization and functional identity of the sparse visual vocabulary. (a) Increasing the threshold reduces the mean number of active bits and lowers accuracy. Error bars denote the standard deviation across images. (b) Probability mass function (PMF) and cumulative distribution function (CDF) of the image-wise active-bit count $K(x)$. (c) Rank–frequency curve of visual-token activations with a Zipf fit over ranks 1–1200. (d) The visual vocabulary separates into a high-frequency head, a Zipf-like core, a tail, rare tokens, and dormant positions. (e) Empirical CDFs of Jaccard overlap show a sharp separation between same-class and different-class image pairs. (f) Targeted deletion degrades accuracy more rapidly than random deletion.

cosine learning-rate schedule, and 150 training epochs. The primary model used $\tau = 15$, whereas each threshold in the sparsity sweep was trained as a separate model under the same protocol.

For the semantic-generative probe, the trained ResNet-50 was kept frozen, and its active-bit sets were used to condition SD1.5. Only the 2048×768 visual-token embedding matrix E was optimized with AdamW at a learning rate of 5×10^{-4} for 2 epochs. The VAE, CLIP encoder, and U-Net remained frozen throughout this stage. Additional implementation details are provided in the Appendix.

4.2 Recognition with Sparse Binary Representations

To quantify the recognition–sparsity trade-off, we independently fine-tuned a ResNet-50 for each sampled threshold τ under the same protocol. For each model, we measured the mean active-bit count and validation Top-1 accuracy. Figure 2a summarizes the resulting trade-off, with error bars on the active-bit curve denoting the standard deviation across images. Increasing τ reduced the mean active-bit count, with progressively smaller reductions at higher thresholds. For $\tau < 15$, accuracy fluctuated only slightly, whereas it declined clearly once $\tau > 15$. At $\tau = 15$, the model achieved 78.54% Top-1 accuracy while activating only 24.63 of the 2048 bits per image on average. This result demonstrates that substantial recognition performance can be retained with only a few dozen active bits per image. We therefore selected $\tau = 15$ for the subsequent analyses.

4.3 Emergent Structure of the Sparse Visual Vocabulary

Adaptive sparse image descriptions. The binary bottleneck represented each image as an active-bit set $\mathcal{S}(x)$ drawn from the shared visual vocabulary of 2048 tokens. Across the complete ImageNet training split at $\tau = 15$, $K(x)$ had a mean of 24.63 and a standard deviation of 6.88, with a median of 24 and a range of 4–88. Only 1.20% of visual-token positions were active per image on average, corresponding to 98.80% sparsity. The probability mass function (PMF) and cumulative distribution function (CDF) in fig. 2b show that this sparsity was pervasive, with 99.99% of images activating no more than 66 bits. Of the 2048 positions, 117 were never activated. Across the remaining 1931

positions, the mean binary entropy was 0.088 bits per token, and the summed marginal entropy was approximately 170.2 bits. These statistics show that the bottleneck combines extreme per-image sparsity with substantial information capacity, enabling compact yet expressive visual descriptions.

Hierarchical visual-token reuse. To characterize reuse across images, we ranked visual tokens by their activation count over the complete training split. The resulting rank–frequency curve in fig. 2c exhibited a broad Zipf-like core, a distinct high-frequency head, and a steep low-frequency tail. A Zipf fit, $f(r) = Cr^{-\alpha}$, over ranks 1–1200 yielded $\alpha = 0.6589$ and $R^2 = 0.9505$ in log–log space. The first 15 high-frequency tokens represented 0.7% of the visual vocabulary and accounted for 8.2% of all activations, but fell below the fitted relation. The following 57.9% of visual-token positions, spanning ranks 16–1200, formed the broad Zipf-like core and accounted for 83.9% of all activations. Beyond rank 1200, 32.8% of positions formed a steeper tail, were active in more than 10^{-4} of images, and contributed 7.9% of all activations. A further 2.9% of positions were activated at least once but in at most 10^{-4} of images, while the remaining 5.7% were dormant. This head–core–tail profile resembles the multi-regime rank–frequency organization reported in natural-language vocabularies (Ferrer i Cancho & Solé, 2001; Petersen et al., 2012; Piantadosi, 2014). The resulting hierarchy comprises a frequently reused common core and a substantial low-frequency periphery of visual tokens, as summarized in fig. 2d.

Class-structured visual-token combinations. We next tested whether active-bit sets combine shared visual tokens in category-relevant ways. Using class-balanced samples of $500{,}000$ same-class and $500{,}000$ different-class image pairs, we measured the Jaccard overlap of their active-bit sets, namely their intersection-over-union. The mean Jaccard overlap was 0.4983 for same-class pairs and 0.0191 for different-class pairs, yielding a mean gap of $\Delta J = 0.4792$. Class-bootstrap resampling yielded a 95% confidence interval of $[0.4686, 0.4898]$ for this gap. The empirical CDFs in fig. 2e showed a sharp separation, with $J = 0$ for only 0.28% of same-class pairs but 66.67% of different-class pairs. These results reveal a category-relevant combinatorial organization of the visual vocabulary: same-class images recurrently select coherent visual-token combinations, whereas different-class images select substantially different combinations from the same shared vocabulary.

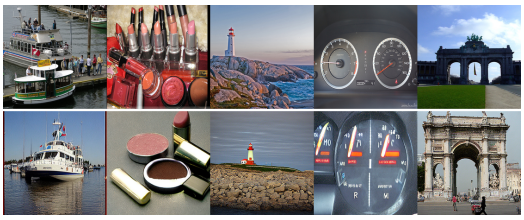
4.4 Counterfactual Robustness and Functional Identity

The discrete binary representation permits direct counterfactual deletion of active visual tokens. For an image with $K(x)$ active tokens, a deletion ratio p removes exactly $n_p(x) = \lfloor pK(x) \rfloor$ tokens under both random and targeted interventions. Under random deletion, we removed $n_p(x)$ randomly selected active tokens to assess robustness to nonspecific token loss. As shown in fig. 2f, accuracy declined gradually despite the extreme sparsity of the representation. At a 20% deletion ratio, random deletion retained 76.98%, only 1.56% below the undeleted baseline. Even after randomly removing half of the active tokens, the model retained 71.03%. These results show that recognition is supported by distributed visual-token contributions rather than by a single indispensable token.

We then tested whether active visual tokens have distinguishable functional roles in the current decision. For the original Top-1 prediction $\widehat{y}(x)$ and runner-up class $y_2(x)$, active token i contributes $I_i(x) = W_{\widehat{y}(x),i} - W_{y_2(x),i}$ to the predicted logit margin. Because the classifier is linear in the binary representation, deleting token i changes this margin by $-I_i(x)$ while all other tokens remain fixed. Targeted deletion therefore removed the $n_p(x)$ active tokens with the largest values of $I_i(x)$. At 20% and 50% deletion, targeted removal reduced accuracy to 56.79% and 14.18%, respectively. Relative to random deletion at the same ratios, targeted removal caused additional losses of 20.19% and 56.85%. This contrast establishes functional differentiation at the classifier level: tokens with the largest analytically specified margin contributions exert disproportionately large counterfactual effects. Together with the reuse and combination analyses, these results characterize the visual vocabulary as a structured operational representation whose active tokens support recognition through distributed yet individually attributable and counterfactually testable functional roles.

4.5 Semantic and Generative Probing of Visual Tokens

Set-conditioned image generation. After learning the visual-token embedding matrix, the active-bit set $\mathcal{S}(x)$ of each image was supplied to Stable Diffusion 1.5 as the sole conditioning signal. Averaged over the validation split, the resulting diffusion noise-prediction loss was 0.1618. Figure 3a compares the source images (top) with their set-conditioned generations (bottom). The gen-



(a) Set-conditioned image generation.



(b) Bit-specific spatial responses.



(c) Single-bit generation.



(d) Two-bit generation.

Figure 3: Semantic and generative probing of learned visual tokens. (a) Five images from ImageNet (top) and samples generated by Stable Diffusion 1.5 (SD1.5) from their corresponding active-bit sets (bottom). (b) The five most strongly activated images from ImageNet for bit 8 (top) and bit 226 (bottom) overlaid with heatmaps. (c) Samples generated only conditioned on bit 8 (top) or bit 226 (bottom) by SD1.5. (d) Samples generated jointly conditioned on bits 8 and 226 by SD1.5.

erated samples recovered semantically recognizable visual content from their source images despite conditioning only on the sparse visual-token set. This result demonstrates that the learned sparse representation can be semantically decoded by a frozen external generative model.

Visual selectivity and lexical readout of non-label concepts. We visualized individual visual tokens by selecting the five ImageNet images with the largest activation for each token and overlaying spatial heatmaps. See the Appendix for the construction of the heatmap. We selected bits 8 and 226 as illustrative cases because their responses concentrated on visual concepts beyond the ImageNet label taxonomy, as shown in fig. 3b. Bit 8 responded selectively to statue-like structures, whereas bit 226 responded to cream-cake-like food. The high-response images for bit 8 carried the labels *pedestal, plinth, footstall* and *monastery*. Those for bit 226 carried the labels *bakery, bakeshop, bakehouse* and *candle, taper, wax light*. Neither *statue* nor *cream cake* is an ImageNet label.

We then compared the corresponding learned visual-token embeddings with the original CLIP token-embedding matrix using cosine similarity. For bit 8, the five nearest CLIP tokens were $\mathcal{N}_5(8)$: $\{seal, statue, statues, column, sculpture\}$. This lexical neighborhood closely matched the statue-like selectivity revealed by visualization. For bit 226, the nearest-token set was $\mathcal{N}_5(226)$: $\{taped, tery, engraved, made, retro\}$. These tokens did not directly name the cream-cake-like content revealed by the response maps. This lexical mismatch may reflect the fact that the phrase *cream cake* is represented compositionally rather than by a single lexical token in the CLIP vocabulary. The learned bit 226 embedding nevertheless provides a single visual-token embedding for this coherent compound visual pattern. These cases show that image-level classification can organize visual tokens around visual regularities beyond the supervised label taxonomy and make them accessible to lexical probing in an external language-grounded embedding space.

Single-token generation. After training, a single visual token could be supplied as the sole active visual condition to the SD1.5 model, providing a direct generative test of its visual effect. We separately conditioned the model on bit 8 and bit 226. As shown in fig. 3c, bit 8 consistently generated statue-like structures across sampled seeds, whereas bit 226 consistently generated cream-cake-like images. These multi-seed samples show that individual visual tokens exert stable semantic effects under the frozen semantic-generative probe. For these cases, a discrete visual token is an independently generatable and inspectable condition with a reproducible visual effect.

The result for bit 226 is particularly informative because its nearest original CLIP tokens do not name its cream-cake-like visual effect. Nevertheless, activating this single token consistently evokes

coherent cream-cake-like images in the frozen generative space. Rather than reproducing the lexical semantics of any single original CLIP token, the learned embedding establishes a new visual conditioning direction within the frozen CLIP–SD1.5 space, enabling coherent cream-cake-like content to be expressed through one visual token. See the Appendix for additional results and analyses.

Generative composition of visual tokens. Having established their individual visual effects, we jointly supplied bits 8 and 226 to the SD1.5 model, with their order randomly shuffled before serialization. As shown in fig. 3d, the resulting samples combined statue-like and cream-cake-like content, including statue-themed cakes, sculptural cake decorations, and cake-like sculptures. Neither visual concept is an ImageNet label, and their joint configuration is outside the observed composition distribution of the ImageNet training distribution used to learn the visual vocabulary. Nevertheless, the frozen generator retained both visual components and produced coherent joint configurations. These observations reveal a general combinatorial capability: an external generator can recombine visual tokens with observed semantic effects, providing an operational bridge from discrete image classification to symbol-like organization.

Emergent symbol-like organization. Together, these experiments show that image-level classification under a sparse binary bottleneck organizes visual information into a shared vocabulary with adaptive sparse image descriptions, hierarchical visual-token reuse, class-structured combinations, and functionally differentiated tokens. The semantic-generative probe further shows that this vocabulary preserves semantic content, supports lexical readout of selected tokens, and enables stable single-token effects and coherent multi-token composition. Beyond the imposed binary discreteness, this organization emerges without language supervision or an explicit symbolic objective, providing an empirical basis for studying how symbol-like structure can arise in neural representations.

5 Discussion

The broader significance of this work lies in a question about language itself: how far can an emergent symbol-like visual representation progress toward language? The learned visual vocabulary exhibits language-associated properties, including sparsity, compositionality, hierarchical reuse, functional differentiation, and stable semantic effects for selected tokens when individually invoked. These findings make the emergence of symbol-like organization an experimentally tractable object of study rather than a purely conceptual possibility. The next challenge is to determine whether this vocabulary can support higher-order linguistic structure beyond token-level semantics and pairwise composition, including grammar-like constraints, syntactic dependencies, sentence-level organization, recursion, and systematic communication. The unordered active set makes this question especially precise because it expresses composition through coactivation rather than a prescribed serial order. It therefore provides a setting for asking whether sequence is fundamental to symbolic organization or simply one modality-specific realization of a more general combinatorial principle. Pursuing these questions could broaden the study of language emergence beyond linguistic agents to perceptual neural systems and invite a broader view of language as one realization of general principles of discrete organization, reuse, and composition.

6 Conclusion

This work shows that visual recognition trained only with image-level supervision can organize a sparse binary representation into a shared visual vocabulary. In an ImageNet-trained ResNet-50, the representation retained 78.54% Top-1 accuracy with only 1.2% of the visual vocabulary active per image on average, while exhibiting hierarchical reuse, category-structured combinations, and functionally differentiated contributions to recognition. A frozen CLIP–Stable Diffusion 1.5 probe further showed that complete active sets generate recognizable visual content, selected tokens produce stable semantic effects, including concepts beyond ImageNet labels, and independently probed tokens can combine into coherent configurations beyond observed ImageNet compositions. At least a subset of the learned visual tokens therefore exhibits multiple properties characteristic of symbolic units, providing a tractable starting point for studying how perceptual symbolic organization may develop toward stable reference, more productive composition, and structured communication.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the grammar, wording, clarity, and readability of the manuscript. We did not use generative AI tools for research ideation, hypothesis formulation, methodology or experiment design, implementation, or the generation of figures or other research artifacts. All AI-assisted text was reviewed and revised by the authors. We take full responsibility for the final content of this work, including all text, claims, results, and artifacts.

REPRODUCIBILITY STATEMENT

We provide detailed specifications for reproducing our experiments in the main paper and Appendix. All experiments use the publicly available ImageNet-1K dataset and the ResNet-50 and Stable Diffusion 1.5 models. We will publicly release the source code implementing the proposed method and experimental procedures to facilitate reproduction and further investigation.

REFERENCES

Subutai Ahmad and Jeff Hawkins. Properties of sparse distributed representations and their application to hierarchical temporal memory. *CoRR*, abs/1503.07469, 2015.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3319–3327. IEEE, 2017.

Brendon Boldt and David R. Mortensen. A review of the applications of deep learning-based emergent communication. *Trans. Mach. Learn. Res.*, 2024, 2024.

Ramon Ferrer I Cancho and Ricard V Solé. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791, 2003.

Luca Colombo, Fabrizio Pittorino, Daniele Zambon, Carlo Baldassi, Manuel Roveri, and Cesare Alippi. Bep: A binary error propagation algorithm for binary neural networks training. In *International Conference on Learning Representations*, volume 2026, pp. 25044–25061, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Zechun Deng, Xiaoming Li, Ziqian Bi, Yiming Lei, Jiawei Xu, Tianyang Wang, Chia Xin Liang, Junhao Song, and Zhenyu Yu. Multi-agent systems and social intelligence: A comprehensive survey on collaboration, communication, and emergent behaviors. *Communication, and Emergent Behaviors*, 2025.

Ramon Ferrer i Cancho and Ricard V Solé. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, 8(3):165–173, 2001.

Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pp. 160–187. PMLR, 2024.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. *Advances in neural information processing systems*, 29, 2016.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Matthew Kowal, Richard P Wildes, and Konstantinos G Derpanis. Visual concept connectome (vcc): Open world concept discovery and their interlayer connections in deep models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10895–10905. IEEE, 2024.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *CoRR*, abs/1604.00289, 2016.

Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era. *CoRR*, abs/2006.02419, 2020.

Jae Hee Lee, Sergio Lanza, and Stefan Wermter. From neural activations to concepts: A survey on explaining concepts in neural networks. *Neurosymbolic Artificial Intelligence*, 1:NAI–240743, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Ruichen Ma, Guanchao Qiao, Yian Liu, Liwei Meng, Ning Ning, Yang Liu, and Shaogang Hu. A&b bnn: Add&bit-operation-only hardware-friendly binary neural network. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5704–5713. IEEE, 2024.

Ruichen Ma, Liwei Meng, Guanchao Qiao, Ning Ning, Yang Liu, and Shaogang Hu. I2e: Real-time image-to-event conversion for high-performance spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1982–1990, 2026a.

Ruichen Ma, Xiaoyang Zhang, Jian Bai, Guanchao Qiao, Liwei Meng, Ning Ning, Yang Liu, and Shaogang Hu. Intrinsically stable spiking neural networks: Overcoming the performance barrier in the absence of batch normalization. *arXiv preprint arXiv:2606.31695*, 2026b.

TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *International Conference on Learning Representations*, volume 2024, pp. 51772–51783, 2024.

Jesse Mu and Noah Goodman. Emergent communication of generalizations. *Advances in neural information processing systems*, 34:17994–18007, 2021.

Andrew Nam, Eric Elmoznino, Nikolay Malkin, James McClelland, Yoshua Bengio, and Guillaume Lajoie. Discrete, compositional, and symbolic representations through attractor dynamics. *CoRR*, abs/2310.01807, 2023.

Allen Newell and Herbert A Simon. Computer science as empirical inquiry: Symbols and search. In *ACM Turing award lectures*, pp. 1975. 2007.

Tuomas P. Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *ICLR*. OpenReview.net, 2023.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Francis Jeffry Pelletier. Compositionality and concepts—a perspective from formal semantics and philosophy of language. In *Compositionality and concepts in linguistics and psychology*, pp. 31–94. Springer International Publishing Cham, 2017.

Alexander M Petersen, Joel N Tenenbaum, Shlomo Havlin, H Eugene Stanley, and Matjaž Perc. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific reports*, 2(1):943, 2012.

Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, 2014.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *CoRR*, abs/2404.16014, 2024.

Mathieu Rita, Florian Strub, Jean-Bastien Grill, Olivier Pietquin, and Emmanuel Dupoux. On the role of population heterogeneity in emergent communication. In *ICLR*. OpenReview.net, 2022a.

Mathieu Rita, Corentin Tallec, Paul Michel, Jean-Bastien Grill, Olivier Pietquin, Emmanuel Dupoux, and Florian Strub. Emergent communication: Generalization and overfitting in lewis games. *Advances in neural information processing systems*, 35:1389–1404, 2022b.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Jiangrong Shen, Qi Xu, Gang Pan, and Badong Chen. Improving the sparse structure learning of spiking neural networks from the view of compression efficiency. In *International Conference on Learning Representations*, volume 2025, pp. 92834–92846, 2025.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *International conference on machine learning*, pp. 31504–31520. PMLR, 2023.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *CoRR*, abs/2406.06525, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *CoRR*, abs/2605.29358, 2026.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

HF Wang, ZF Wang, and MN Du. Methods for interpreting and understanding deep neural networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA*, pp. 14–19, 2020.

Kashu Yamazaki, Viet-Khoa Vo-Ho, Darshan Bulsara, and Ngan Le. Spiking neural networks and their applications: A review. *Brain sciences*, 12(7):863, 2022.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37: 128940–128966, 2024.

Chunyu Yuan and Sos S Agaian. A comprehensive review of binary neural network: C. yuan, ss agaian. *Artificial Intelligence Review*, 56(11):12949–13013, 2023.

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE transactions on emerging topics in computational intelligence*, 5(5):726–742, 2021.

George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books, 2016.

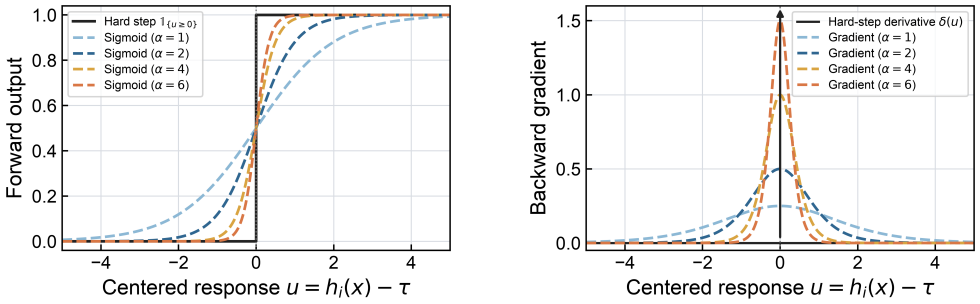
A Additional Methodological Details

A.1 Optimization Through SigmoidSTE

For bit i, let $u_i(x) = h_i(x) - \tau$ denote the feature response centered at the activation threshold. The forward binary representation applies the exact hard threshold $z_i(x) = \mathbf{1}_{\{u_i(x)\geq 0\}}$, which preserves the discrete binary representation used by the classifier. However, the hard threshold has zero ordinary derivative away from $u_i(x) = 0$ and is undefined at the threshold itself.

We therefore optimize the model using a sigmoid straight-through estimator (SigmoidSTE). The forward pass retains the hard binary activation, whereas the backward pass substitutes the derivative of the smooth surrogate $\tilde{z}_i(x) = \sigma(\alpha u_i(x))$. Specifically, the backward gradient is defined as

$$\frac{\partial z_i(x)}{\partial h_i(x)} \coloneqq \frac{\partial \tilde{z}_i(x)}{\partial h_i(x)} = \alpha\, \sigma(\alpha u_i(x))\,[1 - \sigma(\alpha u_i(x))]\,. \tag{14}$$



(a) Forward functions.

(b) Backward gradients.

Figure 4: (a) The forward binary operation remains a hard step, while sigmoid functions illustrate surrogate families with different sharpness values. (b) The hard-step derivative is a Dirac impulse at the threshold, whereas sigmoid derivatives provide surrogate gradients for backpropagation.

Table 1: Definitions and usage shares of the visual-vocabulary strata in fig. 2d. Here, r is the activation-frequency rank and $p_i = f_i/N$ is the per-image activation probability.

Criterion	Head	Zipf-like core	Tail	Rare	Dormant
Condition	$1 \leq r \leq 15$	$16 \leq r \leq 1200$	$r > 1200,\ p_i > 10^{-4}$	$0 < p_i \leq 10^{-4}$	$p_i = 0$
Token share	15 (0.7%)	1185 (57.9%)	672 (32.8%)	59 (2.9%)	117 (5.7%)
Activation share	8.2%	83.9%	7.9%	$< 0.01\%$	0%

The sharpness parameter α controls this approximation. Larger values make the surrogate closer to the hard threshold and concentrate the gradient magnitude more tightly around it. We use $\alpha = 4$ throughout the recognition experiments. Figure 4a visualizes the forward surrogate family and fig. 4b visualizes its corresponding backward gradients.

A.2 Capacity-Bounded Serialization of Active-Bit Sets

The CLIP encoder accepts a context of $L = 77$ tokens. To reserve one position for the BOS token and at least one position for EOT, the implementation permits at most 75 visual tokens per image. Specifically, we first obtain the indices $\mathcal{I}_{75}(x)$ of the 75 largest entries of $\mathbf{h}(x)$ and then retain the threshold-exceeding indices: $\widetilde{\mathcal{S}}(x) = \{i \in \mathcal{I}_{75}(x) \mid h_i(x) \geq \tau\}$. The resulting set is randomly shuffled before serialization, as described in the main text. Thus, the top-75 operation only constructs a capacity-bounded active-bit set. Its response-based ordering is not passed to the CLIP encoder.

For all images with $K(x) \leq 75$, $\widetilde{\mathcal{S}}(x) = \mathcal{S}(x)$. Only 20 ImageNet training images had more than 75 active bits, corresponding to approximately 1.56×10^{-5} of the training split. For these rare cases, the implementation retains the 75 strongest threshold-exceeding responses before random serialization. This procedure prevents context-length overflow while preserving the unordered set-based conditioning protocol used throughout the semantic-generative probe.

A.3 Usage-Stratum Definitions

Table 1 defines the descriptive frequency strata used in fig. 2d.

A.4 Jaccard Pair Sampling

The Jaccard analysis used the ImageNet-1K validation split. For the same-class condition, we randomly sampled 500 pairs of distinct images from each class, for a total of 500,000 pairs. For the different-class condition, each class served as an anchor. We sampled 500 distinct comparison classes without replacement from the remaining 999 classes. For each selected comparison class, one image was randomly sampled from the anchor class and one from the comparison class, yielding 500,000 different-class pairs in total.

Table 2: Recognition, active-set, and visual-vocabulary statistics across the threshold sweep. Top-1 accuracy is measured on the ImageNet validation split. $K_{99.99}$ denotes the 99.99% quantile of the image-wise active-bit count, N_{act} is the number of tokens activated at least once, and $H_{\text{marg}} = \sum_i h_2(p_i)$ is the total marginal entropy. Zipf fits use ranks 1–1200.

	Recognition	Image-wise activity		Vocabulary utilization		Rank–frequency fit	
τ	Top-1 (%)	$\mathbb{E}[K] \pm$ s.d.	$K_{99.99}$	N_{act}	H_{marg} (bits)	α	R^2
4	79.48	67.98 ± 23.94	217	2048	407.71	0.4076	0.9609
5	79.45	56.89 ± 18.57	178	2048	352.52	0.4423	0.9594
6	79.43	50.39 ± 16.04	154	2048	318.15	0.4763	0.9531
7	79.31	44.93 ± 14.31	135	2048	289.49	0.4982	0.9513
8	79.40	40.42 ± 12.46	118	2048	264.71	0.5227	0.9534
9	79.25	36.72 ± 10.97	105	2045	243.82	0.5472	0.9522
10	79.12	33.83 ± 10.06	96	2040	227.16	0.5640	0.9538
11	79.06	31.29 ± 9.29	88	2032	211.97	0.5857	0.9525
12	79.09	28.65 ± 8.20	79	2010	195.99	0.6066	0.9538
13	78.80	26.82 ± 7.69	73	1988	184.81	0.6268	0.9485
14	78.85	25.28 ± 7.13	66	1959	174.76	0.6456	0.9503
15	78.54	24.63 ± 6.88	66	1931	170.18	0.6589	0.9505
16	78.38	23.74 ± 6.51	61	1901	164.49	0.6697	0.9537
17	78.16	23.06 ± 6.19	58	1870	159.54	0.6778	0.9600
18	77.85	22.42 ± 5.99	57	1848	156.56	0.6647	0.9600
19	77.37	21.72 ± 5.75	53	1827	151.06	0.6699	0.9688
20	76.81	20.91 ± 5.79	51	1663	143.33	0.7222	0.9611

A.5 Random-Deletion Evaluation

At each deletion ratio, the reported random-deletion Top-1 accuracy was averaged over five independently sampled masking runs.

A.6 Spatial Visualization of Individual Visual Bits

To visualize the spatial selectivity of each visual bit, we first rank ImageNet validation images by the corresponding response $h_i(x)$. For visual bit i, the top-k image set is

$$\mathcal{T}_k(i) = \mathrm{TopK}_{x \in \mathcal{D}_{\text{val}}} h_i(x), \tag{15}$$

where $\mathcal{D}_{\text{val}}$ denotes the validation split. Using the continuous response $h_i(x)$, rather than the binary activation $z_i(x)$, enables image ranking even when multiple images share the same binary activation.

For each selected image $x \in \mathcal{T}_k(i)$, we extract the ith channel of the final ResNet feature map, $\mathbf{F}_{4,i}(x) \in \mathbb{R}^{7\times 7}$. Because $h_i(x)$ is obtained by global average pooling over this same channel, $\mathbf{F}_{4,i}(x)$ directly provides the spatial response pattern underlying bit i. We retain only locations whose local response is at least the binary activation threshold:

$$\mathbf{M}_i(x)_{u,v} = \mathbf{F}_{4,i}(x)_{u,v} \mathbf{1}_{\{\mathbf{F}_{4,i}(x)_{u,v} \geq \tau\}}, \qquad (u,v) \in \{1,\dots,7\}^2. \tag{16}$$

Thus, locations below τ are set to zero, and all remaining locations retain their original magnitudes.

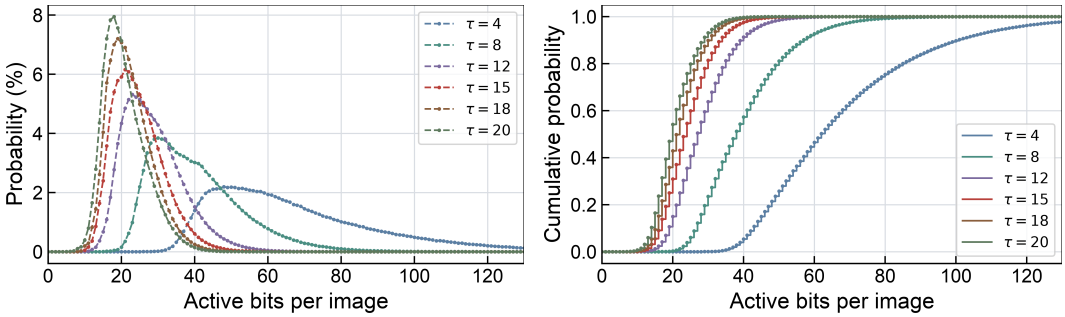
The thresholded map is bicubically upsampled to the 224×224 input resolution, independently normalized to $[0,1]$, converted to a color heatmap, and alpha-blended with the source image using a blending weight of 0.5. The resulting overlays identify spatial regions with strong local responses for each bit, while the top-k image sets reveal the visual patterns associated with its strongest activations.

B Additional Experimental Results and Analyses

B.1 Visual-Vocabulary Structure Across Sparsity Levels

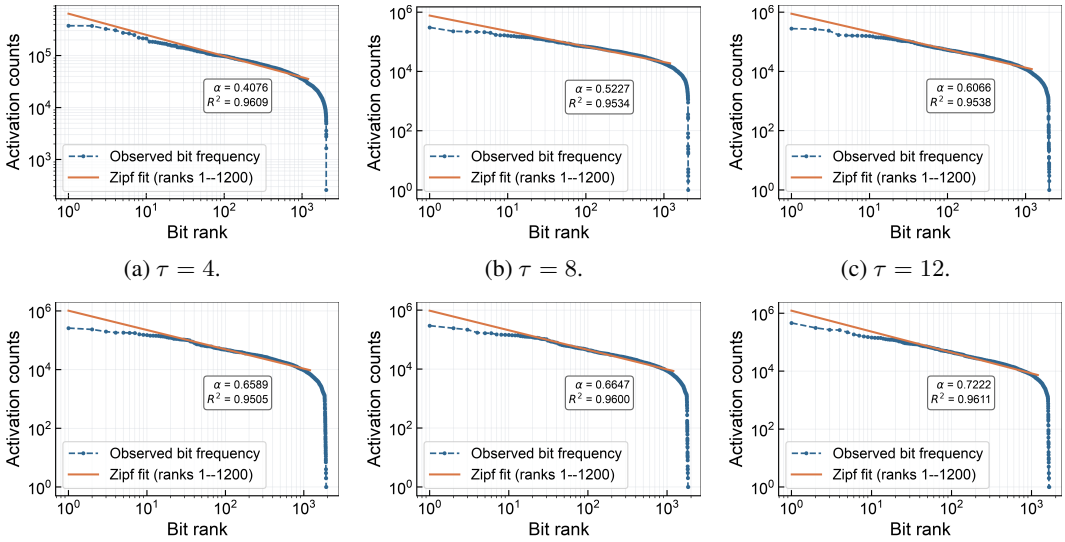
To test whether the visual-vocabulary structure observed at the primary setting persists across sparsity levels, we independently trained a ResNet-50 for each threshold $\tau \in \{4,\dots,20\}$ under the same protocol. For visual bit i, let $p_i = N^{-1} \sum_{x \in \mathcal{D}_{\text{train}}} z_i(x)$ denote its activation probability over the training split. We summarize bit-wise information diversity by the total marginal entropy

$$H_{\text{marg}} = \sum_{i=1}^{2048} h_2(p_i), \qquad h_2(p) = -p \log_2 p - (1-p)\log_2(1-p). \tag{17}$$



(a) Probability mass functions. (b) Cumulative distribution functions.

Figure 5: Threshold-dependent active-set distributions. (a) Probability mass functions and (b) cumulative distribution functions of the image-wise active-bit count $K(x)$ on the ImageNet independently trained models. Increasing τ shifts the distributions toward fewer active bits and reduces the spread.



(a) $\tau = 4$. (b) $\tau = 8$. (c) $\tau = 12$.

(d) $\tau = 15$. (e) $\tau = 18$. (f) $\tau = 20$.

Figure 6: Threshold-dependent rank–frequency structure. Each panel shows the rank–frequency curve of non-dormant visual-token activation counts for one independently trained model. Blue curves show activation counts, and orange lines show Zipf fits, $f_{(r)} = Cr^{-\alpha}$, over ranks 1–1200. A broad Zipf-like regime persists across the sweep, while the fitted exponent increases overall with τ.

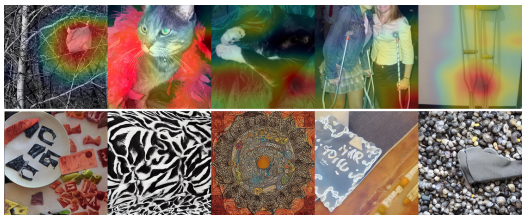
Table 2 and figs. 5 and 6 summarize the resulting recognition, active-set, and vocabulary statistics.

As τ increased, the representations became progressively sparser: the mean active-bit count decreased from 67.98 at $\tau = 4$ to 20.91 at $\tau = 20$, while its standard deviation decreased from 23.94 to 5.79. Accordingly, the active-set distributions shifted toward smaller and more concentrated image-wise active sets. Nevertheless, vocabulary utilization remained broad across the sweep: even at $\tau = 20$, 1663 visual tokens were activated at least once over the complete training split.

The rank–frequency curves retained a similar head–core–tail structure throughout the sweep. The fitted exponent α increased overall from 0.4076 to 0.7222 as τ increased from 4 to 20, indicating a progressively steeper hierarchy of visual-token reuse as the representation became sparser. A natural interpretation is that higher thresholds preferentially preserve activation events in the stronger and more recurrent tails of token-response distributions. When tokens differ in how frequently they produce high responses across images, this selection amplifies their frequency differences: broadly reused tokens remain active across many images, whereas tokens associated with more specific



(a) Token 1846: highest activation frequency.



(b) Token 1672: second-highest activation frequency.



(c) Token 1562: largest lexical-proximity score.



(d) Token 625: second-largest lexical-proximity score.



(e) Token 28: no direct lexical alignment.



(f) Token 149: no direct lexical alignment.

Figure 7: Additional lexical and generative probes of individual visual tokens. Each panel analyzes one visual token. The top row shows the five ImageNet validation images with the strongest responses, overlaid with spatial heatmaps, whereas the bottom row shows five SD1.5 samples generated from the corresponding token alone across random seeds. (a,b) The most frequently activated tokens exhibit broad appearance-level response patterns. (c,d) The tokens with the largest lexical-proximity scores show concordant response maps and generations for elephants and space shuttles. (e,f) Tokens without a semantically matching original CLIP neighbor nevertheless produce stable visual effects corresponding to muzzled animals and sailboats.

visual conditions become relatively rarer. Thus, increased sparsity sharpens the reuse hierarchy while preserving the overall rank–frequency organization.

B.2 Classifier-Head Structure and Bit–Class Associations

Because the classifier receives only the binary representation $\mathbf{z}(x)$, its weight matrix $W \in \mathbb{R}^{1000\times 2048}$ provides a direct analytical interface between visual bits and ImageNet classes. Holding all other bits fixed, changing bit i from inactive to active changes the class-c logit by $W_{c,i}$. Thus, positive and negative weights respectively raise and lower the class-c logit, with their magnitudes specifying the corresponding logit-scale effects. The image-specific margin score $I_i(x)$ used for targeted deletion is the difference between two such weights.

The learned classifier head exhibited an asymmetric signed structure. Positive weights accounted for 30.26% of all entries and had a mean value of 0.04134, whereas the remaining 69.74% were negative, with a mean absolute value of 0.01794. Thus, a bit generally provides stronger positive support for fewer classes while exerting weaker negative effects on more classes. Each visual bit had positive associations with 302.62 ± 86.75 classes on average (range: 120–677), while each class received positive contributions from 619.77 ± 77.89 bits (range: 399–890).

These signed weights allow the class-level contribution profile of each bit to be read out by ranking $W_{c,i}$ across classes. We use bits 8 and 226, which were examined in the spatial and generative analyses, as illustrative examples. For bit 8, the five largest positive weights were associated with

Table 3: Top-five original CLIP token neighbors for selected visual tokens. Each entry gives the decoded token and its cosine similarity to the learned visual-token embedding. The CLIP end-of-word suffix is omitted, and † denotes a non-word-final BPE fragment.

Token	Nearest 1	Nearest 2	Nearest 3	Nearest 4	Nearest 5
1846	`fantaken` 0.200	`antioxid`† 0.197	`printable` 0.192	`,...` 0.191	`travelblog` 0.190
1672	✓ 0.168	`/@` 0.166	`embarrassing` 0.164	`expat` 0.163	☀ 0.158
1562	`elephant` 0.450	`elephants` 0.377	`eleph`† 0.276	`ellie` 0.195	🐘 0.182
625	`shuttle` 0.442	`shutt`† 0.252	`rocket` 0.187	`eties` 0.183	`endeavour` 0.180
28	😷 0.152	💯 0.142	`annu`† 0.141	`mcel`† 0.137	✕ 0.137
149	− 0.185	— 0.168	—> 0.166	`!`† 0.161	`rc` 0.159
8	`seal` 0.187	`statue` 0.187	`statues` 0.182	`column` 0.168	`sculpture` 0.157
226	`taped` 0.146	`tery` 0.143	`engraved` 0.143	`made` 0.140	`retro`† 0.139

pedestal, plinth, footstall (1.536), *triumphal arch* (0.935), *stupa, tope* (0.911), *totem pole* (0.877), and *fountain* (0.834), forming a group of monument- and structure-related classes consistent with its statue-like response maps and generated images. For bit 226, the corresponding classes were *bakery, bakeshop, bakehouse* (1.095), *candle, taper, wax light* (0.764), *ice cream, icecream* (0.669), *bagel, beigel* (0.649), and *pretzel* (0.603), forming a food- and bakery-related profile consistent with its cream-cake-like visual effect. Together with spatial visualization and single-bit generation, these class-level profiles provide complementary evidence that selected bits have distinct and interpretable functional roles in the learned recognition system.

B.3 Lexical Alignment and Generative Readout of Additional Visual Tokens

We extended the single-token analysis to characterize the lexical alignment of learned visual-token embeddings with the original CLIP token embeddings and their semantic behavior under the frozen semantic-generative probe. For token i, we summarize lexical proximity by $s_i = \max_{j \in \mathcal{V}_{\text{CLIP}}} S_{ij}$, namely its largest cosine similarity to an original CLIP token embedding. Across all 2048 tokens, s_i had a mean of 0.1743 ± 0.0341, a median of 0.1661, and a 5th–95th percentile range of $[0.1431, 0.2404]$. Lexical proximity showed negligible monotonic association with activation frequency across tokens (Spearman $\rho = 0.012$), indicating that frequent reuse and closeness to a single original CLIP token capture distinct properties of the learned vocabulary. Table 3 reports the top-five original CLIP token neighbors for the six additional cases below, together with tokens 8 and 226 from the main text. For each of the six additional tokens, fig. 7 provides a matched qualitative probe: the top row shows its five highest-response images, overlaid with spatial heatmaps, whereas the bottom row shows five single-token generations sampled across random seeds.

We first examined the two most frequently activated tokens over the ImageNet training split: token 1846, which was active for 260,598 images (20.34%), and token 1672, which was active for 234,868 images (18.33%). As shown in fig. 7(a,b), these tokens appear to respond to broad recurrent appearance-level regularities, potentially including texture-related patterns, across visually diverse images rather than to a single, sharply delimited category. Their high activation frequencies are therefore consistent with a role as broadly reused components of the shared visual vocabulary.

We next examined tokens 1562 and 625, which attained the two largest lexical-proximity scores among all tokens: $s_{1562} = 0.4497$ for `elephant` and $s_{625} = 0.4416$ for `shuttle`. Their response maps selectively highlighted elephants and space shuttles, respectively, while their single-

token generations consistently produced corresponding visual content in fig. 7(c,d). For these cases, spatial selectivity, lexical readout, and behavior under the frozen generative probe support the same visual interpretation.

Tokens 28 and 149 show a distinct pattern. Their nearest original CLIP entries were dominated by nonlexical tokens or BPE fragments and did not provide a direct semantic match, yet their response maps and single-token generations consistently reveal muzzled animals and sailboats, respectively, as shown in fig. 7(e,f). Together with token 226 in the main text, these cases show that direct alignment to an individual original CLIP token is not required for a learned visual token to acquire a stable semantic effect. Rather than merely reproducing a pre-existing CLIP lexical direction, the learned embeddings can establish stable visual conditioning directions within the frozen CLIP–SD1.5 space. They thereby allow individual visual tokens to express concepts that lack a direct one-token counterpart in the original CLIP vocabulary.

B.4 Scope and Extensibility of the Probing Framework

The probing framework is defined for every visual token in the learned vocabulary rather than only for the illustrative cases shown above. Given a trained recognition model, any token can be characterized using the same operational analyses: its activation frequency, highest-response images and spatial response maps, classifier-head associations, nearest original CLIP tokens, and single-token generations. The selected examples therefore illustrate distinct regimes within the vocabulary, rather than defining a token-specific analysis procedure. Not every token is expected to admit a simple human-readable interpretation, particularly broadly reused tokens whose responses may reflect distributed appearance-level regularities.

The framework is also not inherently restricted to the final ResNet feature vector. An intermediate feature layer can likewise be binarized to define a layer-specific discrete vocabulary, after which a separate visual-token embedding matrix can be optimized through the same frozen semantic-generative probe. Such an extension would permit comparisons of symbolic organization across representational depth. Whether intermediate-layer vocabularies exhibit comparable lexical alignment, generative stability, or compositional behavior remains an empirical question for future work.

C Computational Cost

The primary experiments were conducted on the NVIDIA GeForce RTX 5090. Training the ResNet-50 recognition model required approximately 30 GPU-hours, whereas optimizing the 2048×768 visual-token embedding matrix with SD1.5 for two epochs required approximately 20 GPU-hours. Thus, reproducing the primary results required approximately 50 GPU-hours on this hardware. This estimate excludes the threshold sweep, auxiliary analyses, and exploratory trial runs.